May 2019

A Simulation Platform for Generation of Synthetic Videos for Human Activity Recognition

Gary Plunkett
Western Washington University

Follow this and additional works at: https://cedar.wwu.edu/scholwk
Part of the Higher Education Commons


This Event is brought to you for free and open access by the Conferences and Events at Western CEDAR. It has been accepted for inclusion in Scholars Week by an authorized administrator of Western CEDAR. For more information, please contact westerncedar@wwu.edu.
Objective
Create a simulation platform for outputting synthetic video of common household actions and workflows. Video is intended to be used to train neural network-based action recognition models. Procedural generation and randomness are introduced to increase video generalizability.

Motivation
Labelling video data is expensive!
- Modern computer vision systems are trained on large data sets of real video.
- Need to draw object boundaries on each frame. Known as semantic segmentation (fig 2).

By generating our own synthetic video data, we get labels for free.
- Automatically output semantic segmentation, and other useful ground truth output (fig 3).
- Training with synthetic and real video has been shown to improve performance of computer vision system. [1]

Implementation
Workflow Specification: Manages the action execution order, enforcing conditional and branching execution paths. Interactable objects are specified in the abstract.

Simulation Generator: Translates workflow specification to specific object instances. Introduces variation in lighting, camera angle, object locations, object type, room layout, human model, and agent animation parameters.

Interaction System: Backend system for controlling interactions between agent and household objects. Handles change in game state, object animations, and inventory.

Inverse Kinematics: Inverse kinematics are used to create plausible agent-object interaction animations. The IK system generates an interaction animation based on the agent’s current position, and interaction hand endpoint.

Video Data Pipeline: Video data from multiple modalities is saved, along with synthetic data’s usability by training action recognition systems on one action at a time.

Evaluation
Our simulation platform is designed to facilitate research into workflow recognition systems, where a workflow consists of multiple atomic actions. Before we can investigate workflow recognition, we must validate our synthetic data’s usability by training action recognition systems on one action at a time.

Model: We will use a Multi-Stream Inflated 3D-ConvNet (I3D), an action recognition neural network (fig 4). This model reports state-of-the-art results when trained on multiple inputs streams. [2]

Video Data: Our initial classification task will be kept as simple as possible. The model will be trained to distinguish between a “PickUp” action, and an “Open” action. The streams of video input to be shown in figure 3, with the addition of Instance Segmentation data as well.

Future Work
Workflow recognition, the identification of a string of atomic actions, is an almost completely unexplored area in machine learning. A purely linguistic approach could be layered on top an action recognition network, or multiple recognition networks could operate at different frequencies a la SampleRNN. [3]

Expansion of the simulation platform is another large priority. Currently, all actions are confined to the kitchen and the living room of a home. Adding interactable functionality to the rest of the house, or adding a new home altogether, would increase the diversity of generated video.